Detectando de Câncer de Mama por meio características anotadas e imagens

JULIANA RESPLANDE SANT'ANNA GOMES [1]

[11]UFG – Universidade Federal Goiás
INF – Instituto de Informática
CEP 74.690-900 Goiânia (GO)
julianarsg13@gmail.com

Resumo: Este artigo analisa a detecção de câncer mama utilizando as Redes Perceptron de Múltiplas Camadas (MLP) usando busca em grade com dados provindos de biopsias de tumores de duas formas distintas: a partir de características anotadas dos tumores, como raio e textura do conjunto de dados Breast Cancer Wisconsin, e a partir das imagens dos tumores em si do conjunto de dados Breast Cancer Histopathological (BreakHis). Foram obtidas respectivamente 98% de acurácia para Breast Cancer Wisconsin e 83% de acurácia para o BreakHis.

Palavras Chaves: MLP, câncer de mama, classificação, Breast Cancer Wisconsin, BreakHis.

1 Introdução

Segundo o Instituto Nacional do Câncer (INCA), o câncer da mama é a primeira causa de mortes em mulheres no Brasil, em que muitas vezes o diagnóstico é detectado em uma fase tardia da doença, por conta de uma política ineficaz na detecção da doença [1]. No cenário mundial, o câncer de mama é a segunda categoria de câncer com maior número de mortos, em que o primeiro lugar é o câncer de pulmão, representando um grande problema de saúde pública em todo o mundo [2].

O câncer de mama pode ser detectado através de biópsias, em que se colectam células de tumores para examinar no microscópio [3]. Características dos tumores podem ser anotadas por especialistas para auxiliar na classificação do tumor em benigno ou em maligno, em que neste último caso se caracteriza um quadro de câncer de mama.

O objetivo deste trabalho é comparar a classificação usando MLP com duas categorias distintas de conjuntos dados: um formado características anotadas de imagens (Breast Cancer Wisconsin) e o outro formado pela imagem em si (BreakHis).

2 Metodologia

Para o presente trabalho, as Redes Perceptron de Múltiplas Camadas (MLP) foram empregadas. Como o nome indica, são várias redes Perceptron conectadas. A Equação 1 abaixo mostra a função de ativação de uma rede Perceptron, em que x_j indica a j-ézima entrada x para a rede, wj, o peso associado a entrada x_j, b seria o bias e o φ a função de ativação.

$$a = \phi\left(\sum_j w_j x_j + b\right) \quad (1)$$

No treino MLP, é feita a retro-propagação dos pesos em que os pesos são atualizados segundo as derivadas parciais da função de perda.

O framework Scikit-learn[1] foi utilizado para aplicar MLP, em conjunto com a busca em grade para otimizar parâmetros de forma específica para cada método, como será explicado nas subseções 2.1 e 2.2. Serão disponibilizados os códigos dos experimentos.[2]

2.1 Classificação usando características anotadas de imagens

O método de Classificação usando características anotadas de imagens é analisado utilizando o conjunto de dados Breast Cancer Wisconsin (diagnostic), formado por 569 exemplos de características anotadas de células de tumores mamários, classificados em benignas, composto por 357 exemplos, e malignas com 212 exemplos.

Para cada amostra são fornecidos a média, o desvio padrão e o pior dos seguintes atributos: raio; textura; perímetro; área; suavidade (variação local em comprimentos de raio); compacidade (perímetro ^ 2 / área - 1,0); concavidade (gravidade das porções côncavas do contorno); pontos côncavos (número de porções côncavas do contorno); simetria; dimensão fractal ("aproximação do litoral" - 1) [4].

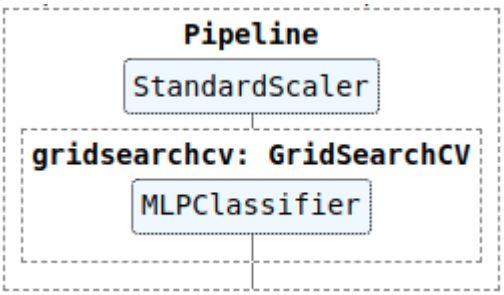

Figura 2.1.1: *Pipeline* de treino

A experimentação foi feita de forma, em que os dados foram divididos em 75% de treino e 15% para teste. A Figura 2.1.1 ilustra o *pipeline* de treino do modelo de classificação. Primeiro, normalizaram-se os valores das *features* por meio do `StandardScaler`, isto é, removendo a média e escalonando para a variância unitária.

Posteriormente, foi feito uma busca em grade (`gridsearchcv`) do classificador MLP (`MLPClassifier`) usando KFOLD com 10 divisões, como ilustra a Figura 2.1.1. Foi e semente 42 para o KFOLD e o MLP; para a busca em grade foram testadas: as funções de ativação φ, como sendo logística, `tanh`, e `relu`; os otimizadores de peso como sendo `sbd` e `adam`; a taxa de aprendizado inicial e alfa entre 1; 0,1 e 0,01. Os parâmetros restantes foram usados os valores padrões fornecidos pela biblioteca Scikit-learn.

2.2 Classificação usando imagens

O método de Classificação usando características anotadas de imagens é analisado utilizando o conjunto de dados Breast Cancer Histopathological Database (BreakHis) com 588 exemplos de tumores benignos e 1242 exemplos de tumores de malignos.

[1] https://scikit-learn.org
[2] https://github.com/jubs12/healthcare-ai

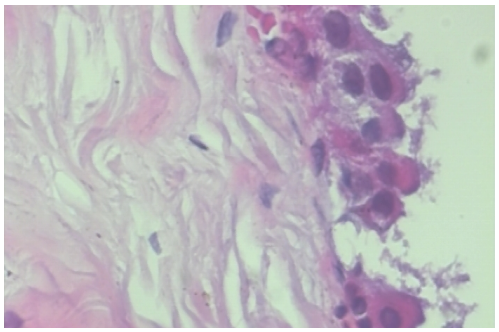

Figura 2.2.1 — Amostra de tumor benigno

As Figuras 2.2.1 e 2.2.2 são amostras do banco dados das classes benigna e maligna respectivamente. No pré-processamento, as imagens tiveram suas dimensões reduzidas de 460x700x3 para 32 x 32 x 3 e foram divididos em 75% de treino e 15% para teste.

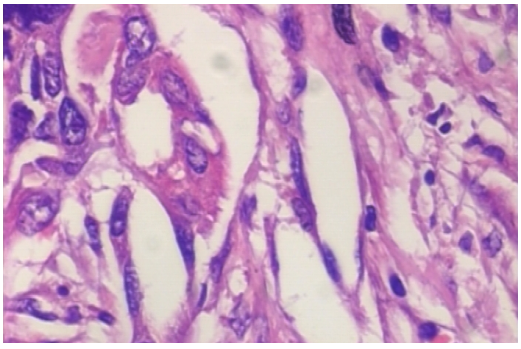

Figura 2.2.1 — Amostra de tumor maligno

Para o treino, foi feito uma busca em grade (`gridsearchcv`) do classificador MLP (`MLPClassifier`), análogo a Figura 2.1.1. O MLP teve como prâmetros alpha igual a 1e-05, otmizador 'adam', número máximo de iterações igual a 1000, batch size 200 e random state 42. A busca em grade, por sua vez, utilizou-se somente acurácia para scoring, mas foram testadas as seguintes dimensões para as camadas ocultas: (100), (100,20), (100,50), (100, 80), (200), (200, 20), (200, 50), (200, 70, 30), (200, 100, 50), (200, 120, 100), (220, 140, 120). Os parâmetros restantes foram mantidos iguais aos fornecidos por padrão pela biblioteca Scikit-learn.

3 Resultados

O melhor modelo de MLP treinado no Breast Cancer Wisconsin obteve acurácia de 99%, possui alfa 1; taxa de aprendizado de 0,1 e otimizador “sgd”; os outros parâmetros são iguais aos valores padrões fornecidos pela biblioteca Scikit-learn. No conjunto de dados BreakHist, o MLP de melhor resultado obteve acurácia de 83%, com a dimensão das camadas ocultas igual a (200, 100, 50). As Tabelas 2 e 3 a seguir ilustram respectivamente os resultados de avaliação por classe em termos de precisão, revocação e F1 para o modelo treinado no Breast Cancer Wisconsin (diagnostic) e para o modelo treinado no BreakHist.

Tabela 1 —Resultados no Breast Cancer Wisconsin (diagnostic)

	Precisão	Revocação	F1
Maligno	0,98	0,98	0,98
Benigno	0,99	0,99	0,99
Macro	0,99	0,99	0,99
Micro	0,99	0,99	0,98

Tabela 2 — Resultados no BreakHist

	Precisão	Revocação	F1
Maligno	0,85	0,91	0,88
Benigno	0,78	0,67	0,72
Macro	0,82	0,79	0,80
Micro	0,83	0,83	0,83

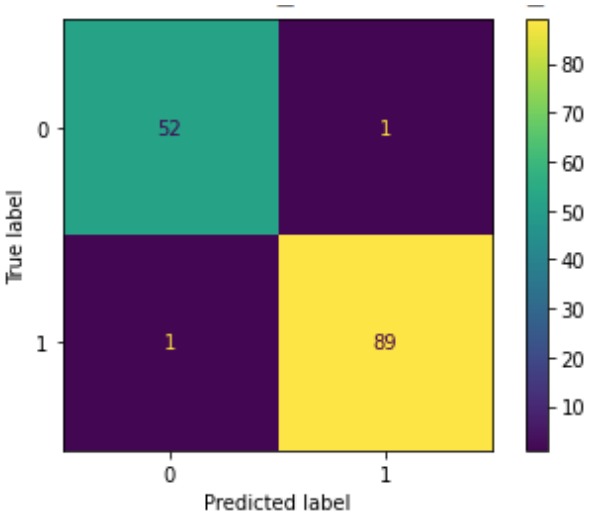

Figura 3.1.1 — Matriz de confusão no Breast Cancer Wisconsin (diagnostic)

As matrizes de confusão são ilustradas nas Figuras 3.1.1 e 3.1.2. No conjunto de dados Breast Cancer Wisconsin (diagnostic), o tumor maligno é representado por 0 e o tumor benigno 1 é indicado por 1. Existe somente um falso positivo e um falso negativo, que afeta mais os resultados das amostras malignas da Tabela 1, uma vez que existem menos exemplos dessa classe.

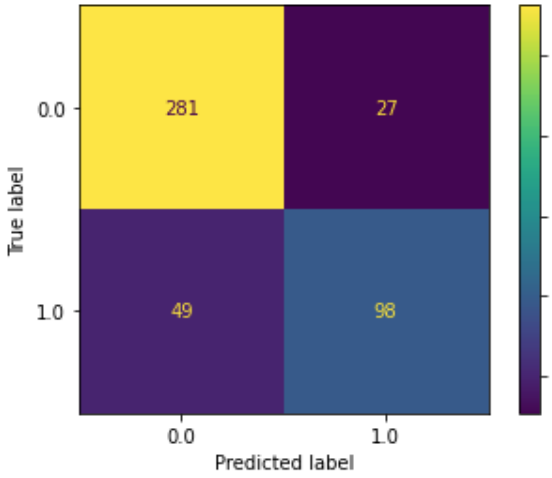

Figura 3.2.1 — Matriz de confusão no BreakHist

4 Conclusões

As redes MLP para a classificação de tumores de biópsias de câncer de mama obtiveram resultados mais satisfatórios com o conjunto de dados Breast Cancer Wisconsin, que possuia as características anotadas, obtendo 98% de acurácia em relação ao conjunto de dados, BreakHist, formados pelas imagens em si com 83% de acurácia.

5 Referências

[1] Silva, Pamella Araújo da, and Sueli da Silva Riul. "Câncer de mama: fatores de risco e detecção precoce." *Revista Brasileira de Enfermagem* 64 (2011): 1016-1021.

[2] Silva, Pamella Araújo da, and Sueli da Silva Riul. "Câncer de mama: fatores de risco e detecção precoce." *Revista Brasileira de Enfermagem* 64 (2011): 1016-1021.

[3] Spanhol, Fabio A., et al. "A dataset for breast cancer histopathological image classification." Ieee transactions on biomedical engineering 63.7 (2015): 1455-1462. .

[4] Scikit-learn. “Breast cancer wisconsin (diagnostic) dataset.” *Scikit-learn*, 2007, https://scikit-learn.org/stable/datasets/toy_dataset.html#breast-cancer-wisconsin-diagnostic-dataset. Accessed 17 10 2021.